DOCUMENT RESUME

ED 318 776 TM 014 865

AUTHOR Webb, Lynn C.; And Others
TITLE Rater Stringency and Consistency in Performance Assessment.
PUB DATE Apr 90
NOTE 28p.; Paper presented at the Annual Meeting of the American Educational Research Association (Boston, MA, April 16-20, 1990) and the Annual Meeting of the National Council on Measurement in Education (Boston, MA, April 17-19, 1990).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Allied Health Occupations; Evaluation Methods; *Evaluators; Higher Education; *Interrater Reliability; Least Squares Statistics; *Licensing Examinations (Professions); *Performance; Regression (Statistics); *Scoring; Test Bias
IDENTIFIERS Oral Examinations; *Rater Reliability; *Rater Stringency Error

ABSTRACT

Two aspects of rater accuracy in performance assessment were analyzed: rater stringency/leniency, and rater consistency. Data were obtained from three administrations of an oral certification examination in a health profession. The examination consists of clinical cases in four content areas or subspecialities. A total of 364 candidates were examined by a total of 60 raters over the 3-year period (1987-89) of the study. In any 1 year, approximately 120 candidates are examined, each by 4 raters; each rater examined from 9 to 14 candidates. Candidates were given three scores by each rater; the scores are on a Likert-type rating scale ranging from 1 to 12. Rater bias was studied using an ordinary least squares approach. Rater consistency was studied by computing each rater's mean-squared deviation from the model. Rater bias could be easily corrected using a simple regression model. Rater inconsistency, when discovered, presented a more complex problem. Inconsistency may be best addressed during the rater training phase. Eight tables present study data. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED318776

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

LYNN C. WEBB

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Rater Stringency and Consistency in Performance Assessment

Lynn C. Webb

Mark R Raymond

Walter M. Houston

American College Testing Program

Paper presented at the annual meeting of the American Educational Research Association and the National Council on Measurment in Education
April 1990, Boston, MA

TM014865

BEST COPY AVAILABLE

ABSTRACT

Data from three administrations of an oral examination in the health professions was analyzed in terms of two aspects of rater accuracy in performance assessment: rater stringency/leniency (bias) and rater consistency. Rater bias was studied using an ordinary least squares approach. Rater consistency was studied by computing each rater's mean-squared deviation from the model.

Rater bias can be easily corrected using a simple regression model. Rater inconsistency, if discovered, presents a more complex problem.

Introduction

Concern about a systematic error component due to differences among raters has received attention in the literature for decades (e.g., Wherry, 1952; Guilford, 1954). Since quantitative ratings in both clinical and nonclinical settings provide one of the most common methods for measurement of competence (Landy & Farr, 1980), the need to acknowledge rater effects is paramount. Oral examinations, essay tests, or any performance assessments that use behavioral ratings scales are in jeopardy of introducing systematic error based on differences among raters that become breaches in standardization. Many procedures to correct for breaches in standardization have become institutionalized in the field of testing: IRT equating, analysis of covariance, and some methods of imputing missing data. Certain performance assessments should be evaluated for standardization based entirely on their format. Specifically, any assessment where candidates are evaluated by a subset of raters is suspect. If all raters evaluated all candidates, any differences in terms of leniency or stringency among raters would not introduce a problem since all candidates would be subjected to the same sources of error. In most cases, the practical considerations of time and expense preclude every rater evaluating every candidate. Far more common is the case of candidates being evaluated by a subset of the total rater group. Random assignment of raters to candidates, large numbers of raters per candidate, and rater training help to decrease the effects of both random and systematic error (Guilford, 1954; Cason & Cason, 1989). However, the effects of systematic error are likely to be significant in applied situations where it is not practical to have candidates rated by large numbers of raters. Certain candidates could be evaluated by a group of the most stringent raters while other candidates were evaluated by the most lenient raters.

The problem of quantifying rater effects in performance assessment data has been approached from a variety of angles. Cason and Cason have pursued the problem of rater effects through a theoretical model attending to the rater's resolving power, sensitivity, stringency and effective rating floor and ceiling effects (1984). Cason and Cason also proposed a linear regression model that addressed rater stringency and, in part, floor and ceiling effects (1985, 1989). This regression approach provided indices of rater bias (leniency, stringency) and estimates of candidate performance that were adjusted or corrected for rater bias.

Cason and Cason (1985) cited common familiarity with regression methods among educational researchers and relatively low computing cost as great advantages to using their approach to minimize lapses in standardization (rater effects), especially in comparison with direct control methods such as enhanced rater training. However, one problem to be noted with this approach is that the regression model as proposed by Cason and Cason cannot be run on commonly available software programs such as SAS or SPSS. A special regression method capable of managing their design is needed.

Another approach to correcting rater effects in performance assessments where raters evaluate a subset of candidates was presented by Wilson (1988). Wilson utilized ordinary least squares regression, followed by weighted least squares to refine the original adjustments to the scores. This addition takes into consideration rater consistency. If the consistency of scoring varies across evaluators, then the regression assumption of equal error variances is violated; consequently, the variances of the parameter estimates will be inflated (Draper & Smith, 1981, p. 110). Furthermore, the parameter estimates (adjusted scores) of candidates evaluated by inconsistent raters will be less accurate than the estimates associated with consistent raters. Wilson's model attempted to address rater inconsistency by deriving parameter weights in a

manner that gives less influence to inconsistent raters. In Wilson's paper, an artificial example with known parameter values, is used to show that the OLS estimates may be very accurate. The subsequent application of weighted least squares added to the accuracy of the estimates. (See de Gruijter, 1984, for a discussion of another least squares approach.)

Two major differences between the approaches taken by Wilson and Cason and Cason are the design matrices and the weighting at the second stage. The design matrix proposed by Wilson can easily be implemented using SAS, SYSTAT, or SPSS. Another difference is that Wilson does not perform the z-transformation of ratings used by Cason and Cason.

The purpose of this study was to analyze two aspects of rater accuracy in performance assessment: rater stringency/leniency and rater consistency. Rater effects for an oral certification examination in the health professions served as the focus.

Methodology

Data Source

Data for this study were obtained from three operational administrations of an oral certification examination administered in a health profession. Raters practice evaluating candidates during a training session with videotaped candidates, and receive feedback regarding their peers' ratings on the same candidates. Raters are scheduled to serve for a number of years, rather than only rating the candidates for one examination administration.

A total of 364 candidates were examined by a total of 60 raters over a three-year period (1987-1989). In any one year, approximately 120 candidates were examined by 40 raters; each candidate was examined by four raters, and each rater examined from 9 to 14 candidates (12 on average). In any given

year, the entire data matrix (40 raters by 120 candidates = 4,800 possible observations) was about 10% complete.

The examination consists of clinical cases in four content areas (four subspecialties). Cases are nested within raters and all candidates receive the same cases. Candidates are given three scores (according to three levels of cognitive processing) by each rater; the scores are on a Likert-type rating scale ranging from 1 to 12. For the purpose of this study, unadjusted sums of the three scores were employed, thus resulting in a score scale with a possible range from 3 to 36 for each rater. A candidate's observed final score is the mean rating obtained over the four raters who evaluated that candidate. This final observed score scale also ranges from 3 to 36 with one-fourth point intervals. It should be noted that the data for this study represents a simplification of the actual scoring procedure.

Procedures

Rating data from the three administrations were subjected to three variations of least squares regression: ordinary least squares (OLS) as described by Wilson (1988); OLS applied to rating data subsequent to subjecting the ratings to a curvilinear transformation (T-OLS) (Cason & Cason, 1986; de Gruijter, 1984); and weighted least squares (WLS) as proposed by Wilson (1988). The three procedures are further described below. All three procedures were applied and compared, with the intent of selecting the single most suitable procedure for detailed analyses.

Ordinary least squares (OLS). A regression method for analyzing incomplete rating data postulates that an observed rating for a candidate is a function of the candidate's true ability and a leniency or stringency effect associated with the rater providing that particular rating. The model also assumes an error component. Consider the following model:

$$y_{ij} = \alpha_i + \beta_j + e_{ij} \qquad i = 1 \text{ to } n,\ j = 1 \text{ to } p$$

where y_{ij} is the score given to candidate i by rater j,

α_i is the true score for candidate i,

β_j is the bias (i.e., leniency) index for rater j

and e_{ij} is random error.

Let a_i denote an estimate of α_i, a candidate's true level of performance. An estimate of β_j, the bias parameter, is denoted by b_j, and corresponds to the magnitude of leniency or stringency error for rater j. This error is referred to as bias throughout this paper, consistent with the notion that the error is systematic as opposed to random. Let k denote the total number of ratings across all candidates and raters. The matrix formulation for the model is:

$$\mathbf{y} = \mathbf{X}\begin{bmatrix}\boldsymbol{\alpha}\\ \boldsymbol{\beta}\end{bmatrix} + \mathbf{e}$$

where $\mathbf{y}$ is a (k x 1) vector of observed ratings,

$\mathbf{X}$ is a (k x (n+p-1)) design matrix,

$\boldsymbol{\alpha}$ is an (n x 1) vector,

$\boldsymbol{\beta}$ is a (p - 1) vector of rater bias indices, and

$\mathbf{e}$ is an (k x 1) vector of random errors.

The design matrix, X, consists of n + p - 1 columns; the column for the last rater is dropped to avoid a linear dependency in the columns of X. Table 1 presents the design matrix for a sample of three candidates and three raters. For all candidates and all raters except for the last rater, the numeral 1 is used to indicate the candidate and rater with which each observed rating is associated; otherwise a zero is used. The ratings associated with the last rater are implied by coding the other p - 1 raters with a minus one (-1). The parameter estimates were then obtained through ordinary least squares regression, where

$$\begin{bmatrix}\mathbf{a}\\ \mathbf{b}\end{bmatrix} = (\mathbf{X}'\mathbf{X})^{-1}\,\mathbf{X}'\mathbf{y}.$$

Transformed ordinary least squares (T-OLS). Cason and Cason (1984, 1986) have recommended that ratings be subjected to a curvilinear transformation to correct for floor and ceiling effects near the end-points of the rating scale. They recommended a probit transformation, which has the effect of stretching the tails of the distribution of ratings. The consequence of the transformation is that observed differences among candidates with extremely high (or extremely low) ratings will be exaggerated. The exaggerated differences compensate for ceiling and floor effects. For the present study, the logit transformation was used. The logit is more commonly used in psychometrics and has more desirable mathematical properties than the probit transformation (Cohen & Cohen, 1983, p. 269).

Weighted least squares (WLS). Weights in the two-stage regression procedure are derived as follows. For each observed rating, a residual is computed to indicate the accuracy with which that rater's observed rating corresponds to the rating predicted by the model. For example, if the value of a_i for candidate i is 3.55 and the value for rater b_j is .75, the model predicts that an observed score, y_{ij} of 4.30 would be assigned by that rater to that candidate. The predicted value of 4.30 is then compared to the observed score, and the difference between the predicted and observed scores are computed and squared. If a rater provided 10 ratings, the mean squared residual for rater j (MSR_j) based on those 10 ratings provides an index of rater consistency. The reciprocals of the mean squared residual ($1/MSR_j$) for all raters are then used as weights in a generalized least square analysis (Draper and Smith, 1981) to obtain revised estimates of the candidates' estimated true scores.

Analysis

The least squares analyses produce a set of beta weights, with the first n weights corresponding to a_i, the adjusted scores (estimated true scores) for each candidate and the last p-1 weights corresponding to b_j (j=1 to p-1) the index of bias for the first p-1 raters. The estimated bias index for the last rater, b_p, is equal to $-\sum_{j=1}^{p-1} b_j$. Three sets of analyses were conducted for each year. First, the OLS, T-OLS, and WLS models were evaluated for each of the three years to assess their degree of similarity. OLS and T-OLS were compared to determine if the logit transformation provided an improved fit of the data to the model. Adjusted ratings provided by OLS were then compared with adjusted ratings provided by WLS to determine if the weighting procedure had a noticeable impact on adjusted ratings.

Second, adjusted and observed ratings were compared to determine the degree of correspondence between the two types of ratings. In addition to correlations, pass/fail decisions were also analyzed to determine if observed ratings and adjusted ratings resulted in the same pass/fail decision for each candidate.

Third, rating behavior was evaluated through a variety of analyses. In addition to identifying the level of rater bias associated with each rater for any given year, bias was examined over a three-year period to determine if raters tended to remain lenient or stringent over time. The tendency to remain lenient or stringent over time is referred to as rater stability. Next, rater consistency was evaluated by computing an index to indicate the degree to which each rater behaved as predicted by the model. The rationale and method for computing this index, MSR, was previously described in the section that explains the weighted least squares procedure. Each rater's indices of consistency were also examined to see if they varied over a three-year period. Finally, specific raters and candidates were singled out for

further study. Specifically, raters who were involved in rating split-decision candidates (i.e., those candidates who would pass by one model, and fail by another) were scrutinized for signs of severe bias and inconsistency.

Results

Comparison of Models

Ordinary least squares (OLS) analysis was applied to observed ratings, as well as to ratings after being subjected to a logistic transformation (T-OLS). The degree to which the data fit the two models was nearly identical for all three years, with the R-squares differing only in the third decimal. The R-squares for OLS were .539, .591, and .525 for 1987, 1988, and 1989, respectively.

The next comparison of interest concerned the fit of the OLS model relative to the fit of the WLS model. The adjusted ratings from the two models were correlated. It appeared that using a two-step weighted model had mixed impact on the results. Although the R-squares for WLS were noticeably higher (by about .05 to .10), the two sets of scores (i.e., WLS and OLS) correlated about .98 on average. Given that the three models behaved so similarly in the present context, the least complex of the models, OLS, was retained for further study at this point in time. Thus, subsequent results are based only on the OLS model.

Characteristics of the Adjusted Ratings

Table 2 presents the descriptive statistics for observed ratings and for ratings adjusted by the OLS method. Note from Table 2 that the observed score means are nearly identical to the OLS means, indicating that candidates were assigned to raters randomly. The most notable result in Table 2 is that the correlations between observed ratings and ratings adjusted by OLS are in the

.90s, suggesting that the rank ordering of the ratings is being affected very little by the OLS model.

To get a closer look at the magnitude of the adjustments, a difference score was obtained by subtracting for each candidate the observed rating from the adjusted rating and then squaring this value. Table 3 presents the relevant descriptive statistics for the three years. The columns of Table 3 corresponding to the minimum- and maximum-difference ratings indicate that the adjustments to the observed ratings range from about +3.00 to -3.00 points, although some scores were adjusted by as much as 4.00 points. The right-most column of Table 3 presents the root-mean-squared differences for each year. As can be seen, candidates' scores are adjusted, on average, by about 1.25 points. Since the observed score scale has a range of about 20 points (from about 12 to 32 points), the adjustments do not appear large. The average adjustment is, in fact, about 6% of the scale ($1.25/20 \approx 6\%$). Another way to gauge the relative magnitude of the adjustment is to compare the average adjustment (1.25) to the standard deviation of the ratings (about 3.80). When evaluated from this perspective, the adjustments are on the order of one-third of a standard deviation. Such adjustments may or may not be significant, depending primarily upon whether the adjustments alter the decisions to be made.

Since these ratings were obtained from a domain-referenced certification examination with important personal and social consequences, perhaps the most meaningful way to investigate the magnitude of the adjustments is to compare the pass/fail decisions for observed and adjusted ratings. For each year's data, several artificial but realistic passing-points were imposed to determine the consistency of pass/fail decisions for each of the passing points, based on observed and adjusted ratings. Seven passing points were selected such that the pass rates would range from about 70% to 90%--pass

rates typical of a certification examination. The level of agreement between pass/fail decisions based on the adjusted versus observed ratings was computed. Coefficient Kappa (Cohen, 1960) was so used as an index of agreement. Kappa is an index of the proportion of consistent classifications beyond that expected by chance, and is frequently used as an index of decision consistency in criterion-referenced testing (Subkoviak, 1980).

Table 4 presents the results of the pass/fail analyses for all three years. The overall pass rate changes very little as a function of rating type (observed versus adjusted). This is not surprising since the distributional properties of the two types of ratings were very similar. However, in each year there were some candidates whose pass/fail decisions would change if adjusted ratings were used. The percentage of candidates for whom the pass/fail decision would change ranges from 3.1% to 10.5%, with an average of about 5.9%. The Kappa coefficients range from .683 to .893, with a mean of .801. The data in Table 4 clearly suggest that, despite the high correlations between adjusted and unadjusted scores, decisions about several candidates (6% of 120 = 7) each year would change if adjusted ratings were used instead of observed ratings.

Examiner Characteristics

The regression coefficient, b_j, for each rater serves as the index of bias. This index is relative, with the constraint the $\sum_{j=1}^{p} b_j = 0$. As with the adjusted ratings, the bias indices are expressed on the same scale as the original ratings. That is, if $b_j = 2.5$, then rater j is 2.5 points too stringent with respect to the original rating scale (which has a possible range from 12 to 36 points). Table 5 presents the descriptive statistics for b_j. As indicated in Table 5, the magnitude of rater biases typically ranges from about -5.0 to +5.0. The standard deviation serves as one index of the average degree of bias. It is, in fact, the square root of the mean squared

deviation from zero bias. To obtain another index of the average magnitude of rater bias, the means of the absolute values of the b_j were obtained for each year. These means appear in Table 5. On the average, raters exhibit more than a ±2 point bias. That is, any rater selected at random would be expected to be lenient or stringent by about two or more raw score points.

Of the 40 raters who serve in any single year, about 30 return the following year. To investigate the extent to which raters tend to remain lenient or stringent over time, the correlations between adjacent years were computed. The correlation between bias indices for 27 raters who served in 1987 and 1988 was .68 ($p < .05$). Similarly, the correlation between bias indices for the 31 raters who served in both 1988 and 1989 was .62 ($p < .05$). Although the correlations demonstrate a moderate to high degree of stability over a one-year period, there were certain raters who showed a high degree of change from one year to the next.[1] For example, from 1987 to 1988 five raters had at least a three-point change in their bias index; from 1988 to 1989, six raters exhibited at least a three-point change.

The degree to which each rater provided ratings as predicted by the OLS model was also evaluated. Table 6 presents the summary statistics for the consistency index. The root-mean-squared residual, $\sqrt{MSR}$, between observed and modelled ratings is, on the average, about 4.0. Although the consistency indices typically range from 2.0 to 7.0, most indices were not too far above or below the mean of 4.0. For any given year, about four to six (10% to 15%)

---

[1] The bias indices are conditional on (i.e., relative to) the knowledge level of the candidates within a specific year and on the leniency or stringency of the particular cohort of raters. Other available data (scores from equated multiple-choice tests) suggest that candidate knowledge is fairly stable over a two-year period. However, bias indices will still be influenced by the composition of a particular cohort. A particular rater's bias may change because the new raters may be more or less biased than the non-returning raters they replaced, or because the level of bias of other returning raters changed. Thus, the absolute comparisons across years are, at best, tenuous.

of the raters had a consistency index above 5.0; while only one or two (2.5% to 5%) had an index that exceeded 6.0. The latter group might be singled out for further study, since there would be reason to doubt the accuracy of both their observed and adjusted ratings. It should be noted that WLS would address these outliers by giving less weight to their ratings.

Consistency indices were also correlated over the two, two-year periods to determine if they tended to remain stable over time. The correlation between consistency indices for 1987 and 1988 was .38 ($n = 27$, $p < .05$); and for 1988 and 1989 the correlation was .28 ($n = 31$, $p < .06$). The magnitude of the correlations suggests that rater consistency is only marginally stable from one year to the next. However, it should also be recognized that the weak relationship may be due, in part, to the lack of stability of the consistency index itself. It is a fairly complex statistic that is a function of several factors.

The final analysis consisted of a subjective review of raters who rated split-decision candidates. These are candidates who would pass if adjusted scores were used and fail if observed ratings were used (and vice versa). This type of analysis may have practical importance since one would be more inclined to trust the adjusted ratings obtained from reasonably consistent raters, and less inclined to trust adjusted ratings, as well as observed ratings, of inconsistent raters. To illustrate the type of information such an analysis might yield, an arbitrary passing point (19.5) for a single year (1987) was selected, and ratings of split-decision candidates and the raters involved were studied.

Candidates' decisions were categorized into a two-by-two matrix, where two of the four cells represent stable decisions across the observed and adjusted ratings and the remaining two cells represent disagreement in decision. The distribution of decisions across cells is shown in Table 7.

Table 7 shows 122 (95%) stable decisions: 21 candidates who failed according to both the observed and adjusted scores, and 101 candidates who passed according to both the observed and adjusted scores. The seven unstable decisions are made up of five candidates who passed according to the observed decision and failed according to the adjusted scores, and two candidates who failed according to the observed decision and passed according to the adjusted scores.

The seven candidates with unstable decisions were evaluated by 21 raters; 14 raters evaluated only one candidate and seven raters evaluated two candidates. Table 8 shows the mean (beta weights) and MSR for the raters associated with the seven unstable decisions from the example. The beta weights and MSR for raters could be considered when deciding between the observed and adjusted scores for unstable decision candidates.

Discussion

Psychometricians are obligated ethically to investigate rater effects when analyzing certification examination data for which every rater has not evaluated every candidate. Results from this study indicate that these investigations need not be elaborate. For these data, neither the T-OLS procedure nor the WLS procedure were found to be superior to the simpler OLS procedure.

Persons for whom this type of study is critical are those whose unadjusted scores fall near the passing score. In most cases, the majority of candidates would not be affected by a decision to ignore rater effects. However, the pass/fail decision may change for candidates near the passing score when some adjustment procedure is implimented.

Failing candidates whose scores are near the passing score are frequently the most vocal in terms of criticizing the test and/or the licensure or

certification board. These candidates have an incentive to closely scrutinize the procedures used to score the test.

Licensure and certification boards must closely scrutinize the purpose of the examination and consider the risks of false-positive and false-negative decisions. All factors, including rater effects, should be considered in order to obtain the best estimate possible of the candidate's true ability.

Licensure and certification boards may be reluctant to correct for rater effects due to the composition of the rater group. It is common for the group of raters to represent the proven leaders in the field, and this may introduce a reluctance to adjust scores assigned by such prestigious evaluators. Leadership in a field, however, does not imply test and measurement expertise. Individual differences are also to be expected, and are clearly evidenced during rater training. Content experts may vary widely on their ratings of the same videotaped simulation of an examination experience. Attempts to train the rater group to a point of complete consensus in rating could be more time-consuming and expensive than a simple statistical correction for rater effects. A balance between rater training and statistical correction could be the most satisfactory alternative for the raters and candidates involved.

Rater inconsistency could be analyzed in two respects: across candidates and over time. It is difficult to pinpoint the causes of inconsistency across candidates. When rater a evaluates candidate b more stringently than the other three raters did, and evaluates candidate c more leniently than the other three raters who evaluated candidate c, what can be deduced? Is rater a evaluating candidates on factors that the other raters are not? Are these factors reflective of the candidate's true ability, or of some extraneous source?

Rater inconsistency over time introduces several possible explanations beyond differing abilities of the candidates. A rater's perspective may

change over time with the additional evaluation experience. Other factors due to raters' experiences from one year to the next could change the aspects of candidate performance that raters focus on, and thus influence the ratings given.

Licensure and certification boards could set policy on the magnitude of adjustment of candidate scores based on the consistency measures of the raters contributing to a candidate's score. Scores that are based on assessments from consistent raters could be adjusted statistically to better reflect the candidates' true ability. When scores are based on assessments from inconsistent raters, the observed scores may be the best estimates of true ability in these cases.

Policy setting on the magnitude of adjustment of scores based on measurements of rater leniency/stringency may be difficult for licensure and certification boards. The concept of differential adjustment of candidates' scores is analogous to other statistical procedures used for many certification examinations. As an example, equating scores from different test administrations to maintain a standard for performance over time is intuitively appealing; candidates who write a more difficult administration than the one on which the standard was set receive points; candidates who write a less difficult administration than the one on which the standard was set lose points. In the case of rater effects, candidates who are evaluated by stringent raters receive points; candidates evaluated by lenient raters lose points. The theory is straightforward and intuitively appealing.

The concept of rater inconsistency is not so easily comprehended. Perhaps the best way to help ensure that raters are consistent is during the training phase. Given an acceptable level of rater consistency, statistical adjustments can be used to correct for rater bias (leniency/stringency) effects. This procedure is analogous to that used in standardized testing.

Test forms are constructed to have similar reliability (consistency). Equating is then used to adjust for differences in difficulty.

Whether or not to ignore rater effects in certification and licensure test data is a policy issue to be decided by certification and licensure boards. The job of psychometricians advising these boards is to clearly delineate the statistical procedures to adjust for rater effects, while considering the ramifications of such procedures.

References

Cason, G.J. & Cason, C.L. (1984). A Deterministic Theory of Clinical Performance Rating. Evaluation and the Health Profession, 7, 221-247.

Cason, G.J. & Cason, C.L. (1985). A Regression Solution to Cason and Cason's Model of Clinical Performance Rating: Easier, Cheaper, Faster. Presented at the annual meeting of the American Educational Research Association, Chicago.

Cason, G.J. & Cason, C.L. (1989). Rater Stringency Error in Performance Rating: A Contrast of Three Models. Presented at the annual meeting of the American Educational Research Association, San Francisco.

Cohen, J. (1960). Coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20(1), 137-146.

Cohen, J. & Cohen, P. (1983). Applied multiple regression/correlation analysis for the behavioral sciences. 2nd ed. Hillsdale, N.J.: L. Erlbaum Associates.

de Gruijter, D.N.M. (1984). Two Simple Models for Rater Effects. Applied Psychological Measurement, 8(2), 213-218.

Draper, N.R. & Smith, H. (1981). Applied regression analysis. 2nd ed. New York: Wiley.

Guilford, J.P. (1954). Psychometric methods. 2nd ed. New York: McGraw-Hill.

Landy, F.J. & Farr, J.L. (1980). Performance Rating. Psychological Bulletin, 87, 72-101.

Subkoviak, M.J. (1980). Decision consistency approaches. In R.A. Berk (Ed.). Criterion-referenced measurement: The state of the art. Baltimore: Johns Hopkins University Press.

Wherry, R.J. (1952). The Control of Bias in Rating: A Theory of Rating (Personnel Research Board Report 922). Washington, D.C.: Department of the Army Personnel Research Section, February.

Wilson, H.G. (1988). Parameter Estimation for Peer Grading Under Incomplete Design. Educational and Psychological Measurement, 48, 69-81.

Table 1

Example of Incomplete Rating Matrix and Corresponding Design Matrix for Three Candidates (C1 to C3) and Three Raters (R1 to R3)

Table 1A
Rating Matrix

	Raters		
Cand.	R1	R2	R3
C1	2	•	4
C2	3	4	•
C3	•	6	5

Table 1B
Design Matrix

Obs. Rating	C1	C2	C3	R1	R2
2	1	0	0	1	0
4	1	0	0	-1	-1
3	0	1	0	1	0
4	0	1	0	0	1
6	0	0	1	0	1
5	0	0	1	-1	-1

Table 2

Descriptive Statistics for Observed Ratings and Ratings Adjusted by OLS

Year	Rating Type	N	Mean	S.D.	Min	Max	Correlation of Obs. & OLS
1987	Obs.	129	22.65	3.78	13.25	31.25	
	OLS	129	22.66	4.03	13.19	32.57	.958
1988	Obs.	114	22.73	3.89	13.25	30.00	
	OLS	114	22.81	3.85	13.19	30.47	.934
1989	Obs.	121	22.29	3.81	11.00	31.50	
	OLS	121	22.29	3.73	11.85	29.89	.947

Table 3

Descriptive Statistics for Difference Ratings

		Difference		Squared Difference	
Year	N	Min	max	Mean	Root-Mean
1987	129	-3.06	3.04	1.32	1.15
1988	114	-2.94	4.14	1.96	1.40
1989	121	-3.91	2.87	1.40	1.18

Table 4

Consistency of Pass/Fail Decisions Based On
Observed and Adjusted Ratings for Selected Pass Points

			Pass Rate		Agreement Indices		
Year	N	Pass Point	Observed Ratings	Adjusted Ratings	Disagree Rate	Agree Rate	*Kappa
1987	129	17.5	89.9%	89.1%	3.9%	96.1%	.791
		18.0	87.6%	86.0%	4.6%	95.4%	.799
		18.5	87.6%	84.5%	3.1%	96.9%	.878
		19.0	83.7%	83.7%	4.7%	45.3%	.828
		19.5	82.2%	79.8%	5.4%	94.6%	.826
		20.0	75.2%	76.7%	6.2%	93.8%	.831
		20.5	72.1%	70.5%	6.2%	93.8%	.849
1988	114	17.5	89.5%	91.2%	3.5%	96.5%	.800
		18.0	86.0%	88.6%	4.4%	95.6%	.802
		18.5	85.1%	85.1%	7.0%	93.0%	.724
		19.0	83.3%	83.3%	8.8%	91.2%	.683
		19.5	79.8%	78.1%	10.5%	89.5%	.685
		20.0	76.3%	77.2%	9.6%	90.4%	.731
		20.5	72.8%	73.7%	7.9%	92.1%	.798
1989	121	17.5	88.4%	88.4%	5.0%	95.0%	.756
		18.0	82.6%	85.1%	4.1%	95.9%	.849
		18.5	80.2%	81.8%	3.3%	96.7%	.893
		19.0	78.5%	78.5%	5.0%	95.0%	.852
		19.5	76.9%	76.0%	5.8%	94.2%	.839
		20.0	76.9%	73.6%	6.6%	93.4%	.823
		20.5	74.4%	69.4%	8.3%	91.7%	.795

*Cohen's (1960) coefficient Kappa, is the proportion of consistent classifications (i.e., level of agreement) beyond that expected of chance).

Table 5

Descriptive Statistics for Rater Bias Indices (b_j) for The Three Years

		Raw Bias				Absolute Bias*	
Year	N	Mean	S.D.	Min	Max	Mean	S.D.
1987	40	0.0	2.57	-4.92	6.08	2.03	1.54
1988	40	0.0	2.90	-4.85	4.87	2.41	1.56
1989	40	0.0	2.63	-4.79	5.81	1.87	1.83

*Since the bias index, by definition, has a mean of zero, a more informative index of the average bias is the mean absolute deviations from zero. The standard deviation of raw bias also serves as a useful index of average bias.

Table 6

Descriptive Statistics for Rater Consistency* Index
for Each of Three Years

Year	N	Mean	S.D.	Min	Max
1987	40	3.00	.92	2.55	7.15
1988	40	3.73	.93	1.34	5.53
1989	40	4.07	1.11	2.15	7.25

*Rater consistency is the root-mean-squared-residual obtained from comparing observed ratings to ratings predicted by the model. (See text.)

Table 7

A Sample Distribution of Decision for Observed and Adjusted Ratings for the Passing Point of 19.5 in 1987

		Adjusted Decision: Fail	Adjusted Decision: Pass
Observed Decision	Fail	21	2
	Pass	5	101

Table 8

Mean Squared Residual (MSR) for the 21 Raters Associated with Unstable Decisions for Seven Candidates

Raters	Number of Candidates with Unstable Decision Rated	Mean
A	2	15.80
B	1	36.31
C	1	27.96
D	1	11.58
J	1	12.92
L	2	12.48
M	1	19.45
N	1	22.41
O	2	21.04
Q	1	13.34
R	1	14.40
T	1	14.88
BB	2	6.71
CC	1	13.15
DD	1	13.13
EE	1	16.13
HH	2	12.31
II	2	12.26
JJ	2	13.22
LL	1	20.17
MM	1	26.60